Львівський національний університет імені Івана Франка

На правах рукопису

Волощук Орест Романович

УДК 004.853+004.855.5

Методи кластеризації на великих масивах даних

01.05.03 — математичне та програмне забезпечення обчислювальних машин і систем

Магістерська робота

Науковий керівник

Годич Олесь Васильович,

кандидат фізико-математичних наук, доцент

Львів — 2011

ЗМІСТ

ВСТУП

Актуальність теми. Сьогодні все частіше виникають задачі, так чи інакше пов'язані із розбиттям масиву об'єктів на групи за певними критеріями. Із розвитком обчислювальної техніки збільшуються також об'єми баз даних, що можуть піддаватись такому аналізу. Виникає необхідність створення оптимальних алгоритмів обробки великих масивів даних.

Мета і завдання дослідження. Метою дослідження є розвиток методики кластеризації великих масивів даних. Для досягнення цієї мети були сформульовані та вирішені такі основні завдання:

— провести детальний аналіз та дослідити ефективність різноманітних алгоритмів кластеризації;

— здійснити контроль якості результатів роботи алгоритмів;

— розробити методи та виявити можливі оптимізації, що дозволять прискорити виконання задачі кластеризації

Об'єкт дослідження. Об'єктом дослідження є методи ієрархічні та плоскі алгоритми кластеризації даних.

РОЗДІЛ 1
ОГЛЯД СТАНУ ПРОБЛЕМИ ТА ОСНОВНІ ПОНЯТТЯ

1.0.1. Історія кластеризації. Термін „кластерний аналіз", вперше використаний Тріоном у [6], означає набір підходів та алгоритмів, призначених для об'єднання схожих об'єктів у групи. Ця технологія знайшла своє застосування в цілій низці галузей наук та є необхідною частиною більшості сучасних засобів аналізу даних.

В 1959 радянський вчений Терентьєв розробив так званий „метод кореляційних плеяд" [8], покликаний здійснювати групування на базі корелюючих ознак об'єктів. Займаючись вивченням кореляцій між різними ознаками озерної жаби, він об'єднав їх в групи за абсолютною величиною коефіцієнту кореляції. Таким чином він отримав дві групи ознак – ознаки із великим та з малим значенням кореляції. Терентьєв назвав ці групи „кореляційними плеядами" та опублікував декілька методів їх аналізу. Це посприяло розвитку методів кластеризації за допомогою графів.

На початку 50х років також вийшли публікації Р. Льюїса, Е. Фікса та Дж. Ходжеса, присвячені ієрархічним алгоритмам кластеризації. Відчутний поштовх технології кластерного аналізу дали роботи Розенблатта про розпізнаючий пристій „перцептрон", котрі поклали початок теорії „розпізнавання без вчителя". Поштовхом до розробки методів кластеризації стала публікація [3] в 1963 році. В своїй роботі автори виходили з того, що для створення ефективних біологічних класифікацій процедура кластеризації повинна використовувати всеможливі показники, що характеризують досліджувані організми, проводити оцінку ступеня схожості між цими організмами, та забезпечувати розташування схожих організмів в одну групу. При цьому сформовані групи повинні бути досить локальними, тобто схо-

жість організмів всередині групи повинна бути більшою, ніж між групами. Подальший аналіз таких груп допоможе вияснити, чи відповідають вони реальним біологічним класифікаціям. Сокел та Сніт, автори цеї роботи, вважали, що виявлення структури розподілу об'єктів у групи допоможе встановити процес утворення цих груп.

1.0.2. Типові задачі. Із кластерів, що повертаються в результаті роботи алгоритмів таксономії, можна виділити типових представників , і надалі працювати не з кожним об'єктом великого масиву, а лише з цими представниками. Це дозволяє суттєво спростити аналіз даних. Кластеризація часто стає складовою якоїсь більшої задачі, інструментом підготовки даних для її розв'язання. Такі задачі завжди пов'язані із пошуком та виділенням змістовних структур із великих масивів даних.

До них можна віднести задачі сегментування та розпізнавання зображень, мовлення, пошуку прихованих закономірностей в даних. На практиці такі задачі виникають при розв'язуванні проблем, що виникають в медицині, соціології, економіці та низці інших сфер діяльності людини. медицині, наприклад, техніка сегментування зображення дозволяє виділяти на томограмах окремі області і на підставі їх форми та забарвлення приймати ставити діагноз. В біології використовуються техніки кластеризації для виявлення взаємопов'язаних груп генів та їх впливу на живі організми. Кластеризація успішно застосовується у маркетингових дослідженнях для виявлення зв'язків між різними групами споживачів та потенційних покупців, цільових аудиторій, та оптимального позиціонування нової продукції. В соціологічних дослідженнях використовується кластеризація даних, отриманих з різних джерел, для спрощення їх подальшого аналізу.

У своїй праці [7] Загоруйко описує одну із таких задач. Новосибірські вчені вивчали причини переселення людей з сіл в міста. Були вислані експедиції в навколишні села, жителям яких задавали приблизно сто анкетних

питань, що стосувались віку, сімейного становища, освіти та ін. Після завершення опитування дослідники постали перед необхідністю аналізувати більш ніж сім тисяч анкет, що містили понад сто питань кожна. Ці дані було введено в програму таксономії, котра повернула сім великих таксонів, середні характеристики яких дозволили дати зібраним даним змістовну інтерпретацію. Наприклад, виділився кластер, що містив переважно жінок середнього віку, котрі мали дорослих дітей в місті. Очевидно, представниці цього таксону, названого дослідниками „бабусі", їхали в місто доглядати за своїми внуками. Решту таксонів опрацьовано аналогічно.

Перед процедурою кластеризації часто дані буває необхідно підготувати. В практиці нормою є випадки, коли для деяких об'єктів бракує частини атрибутів – в такому разі перед кластеризацією необхідно здійснити передбачення цих атрибутів, користуючись наявною інформацією. Також значення атрибутів об'єктів необхідно нормувати.

1.0.3. Основні поняття та означення. Дамо математичне означення кластеризації. Нехай D – множина об'єктів.

Означення 1.1. *Кластеризацією* $C = \{C \mid C \subseteq D\}$ називається таке розбиття D на множини, для якого виконується $\cup_{C_i \in C} = D$ і $\forall C_i, C_j \in C : C_i \cap C_{j \neq i} = \varnothing$. Множини C_i називаються кластерами.

Як вже згадувалось раніше, задача кластеризації полягає в знаходженні такого розбиття C, щоб схожі між собою об'єкти належали до одного кластера, а не схожі - до різних. Необхідно визначити спосіб обчислення схожості об'єктів. Для цього на просторі об'єктів вводиться певна метрика, геометричний зміст якої – відстань між об'єктами. Вона використовується як величина, обернена до міри схожості між об'єктами. На даний момент розроблено і широко застосовується наступний набір метрик:

— евклідова відстань

$$\rho(x, x') = \sqrt{\sum_{i=1}^{n}(x_i - x'_i)^2}$$

— квадрат евклідової відстані

$$\rho(x, x') = \sum_{i=1}^{n}(x_i - x'_i)$$

(використовується для надання більшої ваги об'єктам, розташованим далеко один від одного)

— манхеттенська відстань

$$\rho(x, x') = \sum_{i=1}^{n} \mid x_i - x'_i \mid$$

— відстань Чебишева

$$\rho(x, x') = max(\mid x_i - x'_i \mid)$$

ця метрика дозволяє розрізнити об'єкти, якщо вони відрізняються лише одною координатою

1.0.4. Алгоритми кластеризації. На даний момент значного розвитку набула ціла низка різноманітних алгоритмів кластеризації. Їх можна розділити на групи за різними ознаками:

— спосіб групування

- плоскі
- ієрархічні

— визначеність

- визначені

    - невизначені
- чутливість до форми кластерів
    - кластери мають строго сферичну форму
    - кластери можуть мати довільну форму

Плоскі алгоритми будують одне розбиття вибірки на кластери. На відміну від них, агломеративні будують цілу систему взаємовкладених кластерів. На виході ми отримуємо дерево кластерів, коренем якого служить кластер, що містить усі об'єкти вибірки, а листками є найменші кластери з одним об'єктом кожен.

Визначені алгоритми однозначно ставлять у відповідність кожному об'єкту один кластер. Невизначені не дають такої чіткої інформації. Вони повертають імовірність, із якою об'єкт належить до кожного із кластерів.

На сьогоднішній день широко використовуються наступні алгоритми кластеризації:

- k-means ([5], [1])
- UPGMA ([?])
- DBScan
- FOREL ([7])
- c-means

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. *MacQueen, J. B.* Some methods for classification and analysis of multivariate observations / J. B. MacQueen // *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability.* — 1967. — P. 281–297.
2. *Rosenblatt, F.* The perceptron – a perceiving and recognizing automaton: Tech. Rep. 85-460-1 / F. Rosenblatt: Cornell Aeronautical Laboratory, 1957.
3. *Socal, R.* Principles of Numerical Taxonomy / R. Socal, P. Sneath. — San Francisco: W.H. Freeman, 1963.
4. *Sokal, R.* A statistical method for evaluating systematic relationships / R. Sokal, C. Michener // *University of Kansas Science Bulletin.* — 1958. — Vol. 38. — Pp. 1409–1438.
5. *Steinhaus, H.* Sur la division des corps matériels en parties / H. Steinhaus // *Bulletin of the Polish academy of Sciences.* — 1956. — Vol. 3, no. 4. — Pp. 801–804.
6. *Tryon, R. C.* Cluster Analysis / R. C. Tryon. — Ann Arbor: Edwards Brothers, 1939.
7. *Загоруйко.* Прикладные методы анализа данных и знаний / Загоруйко. — Издательство института математики, Новосибирск, 1999.
8. *Терентьев П. В.* Метод корреляционных плеяд / Терентьев П. В. — Вестник Ленинградского университета, 1959.